(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(51) 。Int. Cl. [7]
G06F 17/27

(11) 공개번호 특2001-0108586
(43) 공개일자 2001년12월08일

(21) 출원번호 10-2000-0029143
(22) 출원일자 2000년05월29일

(71) 출원인 우요섭
인천 남구 도화2동 인천대학교 정보통신공학과 우요섭교수
이수선
인천 남구 도화2동 인천대학교 정보통신공학과 대학원
박현재
인천 남구 도화2동 인천대학교 정보통신공학과

(72) 발명자 우요섭
인천 남구 도화2동 인천대학교 정보통신공학과 우요섭교수
박현재
인천 남구 도화2동 인천대학교 정보통신공학과
이수선
인천 남구 도화2동 인천대학교 정보통신공학과 대학원

심사청구 : 있음

(54) 의미정보를 이용한 이단계 단문 분할 장치

요약

본 발명은 한국어 문장을 용언을 중심으로 정확하게 단문을 효율적으로 추출하는데 목적이 있으며, 이와 같은 목적은 하위범주화 사전의 구문정보와 의미정보를 시소러스의 의미정보와 정합을 하는 수단을 이단계 단문 분할장치에 이용하여 한국어 문장을 단문으로 분할한다. 이단계 단문 분할 장치는 핵심 성분인 보어를 정확히 검출하는 수단과 검출되지 않은 비 핵심 성분을 검출된 보어로부터 의존구조를 확장하여 보다 정확한 단문을 검출하는 수단을 포함하여 구성됨으로써 달성된다. 또한, 이단계 단문 분할을 거친 단문 분할된 결과를 후처리 장치를 거쳐 단문 분할의 정확성을 높인다. 단문 분할에 구문정보 만을 이용하는 방법과는 달리 의미정보를 이용하는 이단계 단문 분할 장치를 이용함으로써 단문과 한국어 문장의 구조 파악이 가능하고 또한 보어의 의미 파악도 가능하다.

대표도
도 1

색인어

단문 분할장치, 하위범주화 사전, 시소러스, 의미 정보, 개념 추출 장치의미추출

명세서

도면의 간단한 설명

도 1은 이단계 단문 분할 장치를 포함한 단문 분할 시스템

도 2은 이단계 단문 분할 장치의 처음 장치인 핵심구조 선택 장치

도 3은 계층적 의미정보를 가지고 있는 시소러스의 구조

도 4은 구문정보와 의미정보를 가지고 있는 하위범주화 사전 구조

도 5은 이단계 단문 분할 장치의 두 번째 장치인 의존구조 확장 장치

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

한국어의 단문 분할은 자연언어 응용시스템의 성능을 좌우하는 중요한 부분이다. 본 발명은 보다 효율적으로 복잡한 한국어 문장을 단문 단위로 분할하기 위해, 한국어 문장의 용언을 중심으로 분할하고 문장의 구조를 파악하는 방법과 파악된 문장의 구조를 이용하여 술어와 보어의 의미 및 관계성을 찾아내는 방법에 관한 것으로, 한국어 문장의 구조 파악과 어휘의 의미를 검출하는 기술로서 일종의 문자 정보처리기술이다.

현재 개발되어 있는 단문 분할장치의 경우, 접속어와 같이 단순한 한국어의 문법적 특징을 적용하는 경우가 있으나 이는 원거리 의존 관계나 관형절 처리 등을 해결할 수 없는 문제가 있고, 특정 영역에 제한적인 의미정보와 조사를 사용하는 경우도 있으나 조사의 변화와 생략되는 경우 문장의 구조 파악을 못하는 문제점과 특정 영역에서 제한적으로 수집된 경험적 자료를 단문 분할의 주요 정보원으로 사용하는 것은 구문적으로 애매성이 있는 복잡한 문장의 경우 구조 파악에 오류가 생기고 따라서 단문의 분할이 힘든 문제점이 발생하게 되는 것이다.

발명이 이루고자 하는 기술적 과제

본 발명의 주된 기술적 과제는 구문 분석과 같은 고비용의 장치를 사용하지 않고도 계산 처리 비용을 적게 들여 단문을 분할하는 장치를 개발하는 것이다. 본 발명은 한국어 문장을 의미정보와 구문정보를 이용하여 단문 분할을 수행함으로써, 조사의 변화와 생략을 포함하는 복잡한 문장의 경우도 용이하게 단문 단위로 분할하고자 하는 것이며, 별도의 의미해석이나 어휘 의미파악 장치 없이도 술어와 보어의 의미 및 관계성을 파악하고자 하는 것이 과제가 된다.

본 발명을 적용하면 한국어 문장을 단문으로 용이하게 분할하여 구문 분석의 성공률을 크게 개선시키고 비용을 절감하게 되며, 또한 한국어 어휘의 의미를 파악함으로써 각종 자연언어 응용 시스템의 성능을 향상을 시키게 되는 것이다. 이와 같은 기술적 과제를 달성하기 위해서 본 발명에서는 이단계 단문 분할 장치를 고안하였으며, 이를 위한 핵심 성분 검출 장치와 의존 구조 확장 장치를 개발하는 것이 본 발명이 이루고자 하는 세부적인 과제가 된다.

발명의 구성 및 작용

도 1은 본 발명의 단문 분할 장치의 전체 시스템을 나타낸 것이다. 한국어 문장을 형태소 분석 장치를 이용하여 형태소 분석을 한 후 이단계 단문 분할 장치를 거쳐 여러 개의 단문으로 분할을 한다. 이단계 단문 분할 장치는 한국어 문장을 용언을 중심으로 하위범주화 사전의 구문정보와 의미정보를 시소러스의 계층적 의미정보와의 정합을 이용하여 문장의 보어를 찾아내는 핵심구조 선택 장치와 찾아낸 보어 이외의 성분을 보어에 포함 및 의존시키는 의존구조 확장 장치로 이루어져 있다. 도 2는 핵심구조 추출 장치를 도시한 것이다. 핵심구조 추출 장치는 문장의 중심어인 용언 즉, 동사, 형용사와 용언화 접사('되다', '하다', '스럽다', '답다')를 용언 추출 장치를 거쳐 찾아낸다. 찾아낸 용언 이외의 명사들은 명사 의미후보 추출 장치를 거쳐 계층적 의미정보를 가지고 있는 시소러스의 의미정보를 추출한다. 하위범주 정보 추출 장치는 용언 추출 장치에서 찾아진 용언의 구문정보와 의미정보를 하위범주와 사전에서 추출한다. 구문정보 비교 장치는 용언을 중심으로 왼쪽에 위치한 명사들의 조사를 하위범주정보 추출 장치에서 추출된 구문정보와 비교한다. 구문정보가 일치된 경우 의미정보 비교 장치에서는 하위범주의 의미정보와 명사의 시소러스의 의미정보를 비교한다. 후보등록 장치는 구문정보 비교 장치와 의미정보 비교 장치를 거쳐 일치된 명사를 보어로 선택하여 후보로 등록한다. 용언의 경우 여러 개의 하위범주정보를 가질 수 있으므로 각각의 하위범주정보들을 구문정보 비교 장치와 의미정보 비교 장치를 거쳐 후보로 등록 장치에 등록한다. 용언의 하위범주정보의 수(k)만큼 반복한 후 후보 등록 장치에 등록된 후보들 중 가장 많은 보어를 가지는 경우를 최적 후보 선택 장치에서 최적 후보로 선택한다. 최적 후보 선택 장치에서 후보의 보어 성분의 개수가 같은 경우 의미정보의 일치 정도가 높은 보어를 가지고 있는 후보를 최적 후보로 선택한다. 도 4은 계층적 의미정보를 가지고 있는 시소러스를 도 3는 구문정보와 의미정보를 가지고 있는 하위범주화 사전의 구조를 나타낸다.

도 5는 의존구조 확장 장치의 구조를 나타낸 것이다. 의존구조 확장 장치는 보어로 선택되지 않은 문장 성분들을 보어와 용언에 포함시키는 장치이다. 먼저, 관형절 처리 장치는 형태소 분석 장치의 결과가 관형절의 형태(용언+[~는, ~ㄴ, ~ㄹ, ~되])인 경우 용언의 오른쪽의 명사를 핵심구조 선택 장치의 구문정보 비교 장치와 의미정보 비교 장치를 거쳐 보어성분으로 포함시킨다. 복합명사 처리 장치는 핵심구조 선택 장치에서 선택된 보어의 왼쪽으로 가장 가까운 명사가 조사를 가지고 있지 않은 경우 보어에 포함시켜 복합명사 사전과 비교하여 복합명사로 등록이 되어 있는 경우 보어로 포함시킨다. 연결 구문 처리 장치는 명사가 ',', '그리고' 등의 접속사를 사이에 두고 나열되어 있는 경우 핵심구조 선택 장치의 구문정보 비교 장치와 의미정보 비교 장치를 거쳐 보어로 포함시킨다. 기타성분 처리 장치는 형태소 분석 장치의 결과가 부사, 관형사, 감탄사, 부호인 경우 가까운 오른쪽 피수식 성분에 포함시키고, 보조사인 경우 왼쪽으로 가장 가까운 용언에 포함시키는 장치이다. 후처리 장치는 핵심구조 선택 장치와 의존구조 파장 장치를 거쳐 나온 단문 분할의 결과가 주어을 가지고 있지 않은 경우 문장의 처음에 주어 성분이 존재하면 의미정보 비교 장치를 거쳐 보어성분으로 추가시키는 장치이다.

위의 각각의 장치들을 수행하는 단문 분할 시스템으로 한국어 문장을 단문으로 분할할 수 있다. 하위범주화 사전의 의미정보와 시소러스의 의미정보를 비교함으로써 의미 단계까지 한국어 문장 분석에 이용하게 되어 보다 정확한 단문 분할이 가능하고, 구문정보 비교 장치와 의미정보 비교 장치를 통과한 명사 및 용언의 경우는 어휘 의미의 자동적 파악은 물론, 하위범주화사전에 의한 자동적인 의미 관계성 추출까지 완료되게 되어, 한국어 의미분석을 위해서도 탁월한 성능을 보인다.

발명의 효과

이상에서 상술한 바와 같이 본 발명은, 한국어 문장에서 용언을 중심으로 하위범주화 사전과 시소러스의 구문정보와 의미정보를 이용하여 단문을 검출하는 방법이다. 기존의 방법의 조사에만 의존한 구문정보를 이용한 방법에서 발생하는 조사의 변화 및 생략에 의한 오 검출 현상을 방지하면서도 특정 자료에 국한되지 않은 효율적으로 단문을 검출하게 된다. 단문 분할장치를 사용하면 한국어의 구문 분석 장치가 단순하게 되어 비용 절감의 효과가 크고, 구문 분석 과정의 모호성을 크게 줄여 고속으로 정확한 구문 분석이 가능하게 된다. 또한 본 발명의 단문 분할 장치는 술어와 보어등의 어휘의 의미 및 의미간 관계성을 자동적으로 파악할 수 있어 한국어 문장의 의미 해석 장치를 상당부분 대체하게 된다. 한

국어 문장의 구문 및 의미의 파악을 가능하게 함으로써 정보검색, 기계번역 등의 다양한 분야의 자연언어 응용 시스템의 품질 향상에 크게 기여할 수 있다.

(57) 청구의 범위

청구항 1.

핵심 구조 선택과 의존 구조 확장을 통한 이단계 단문 분할 장치

의미정보와 구문정보를 이용한 핵심 구조 선택 장치,

각각의 하위범주와 정보에 의해 생성된 후보의 최적 후보 선택 장치,

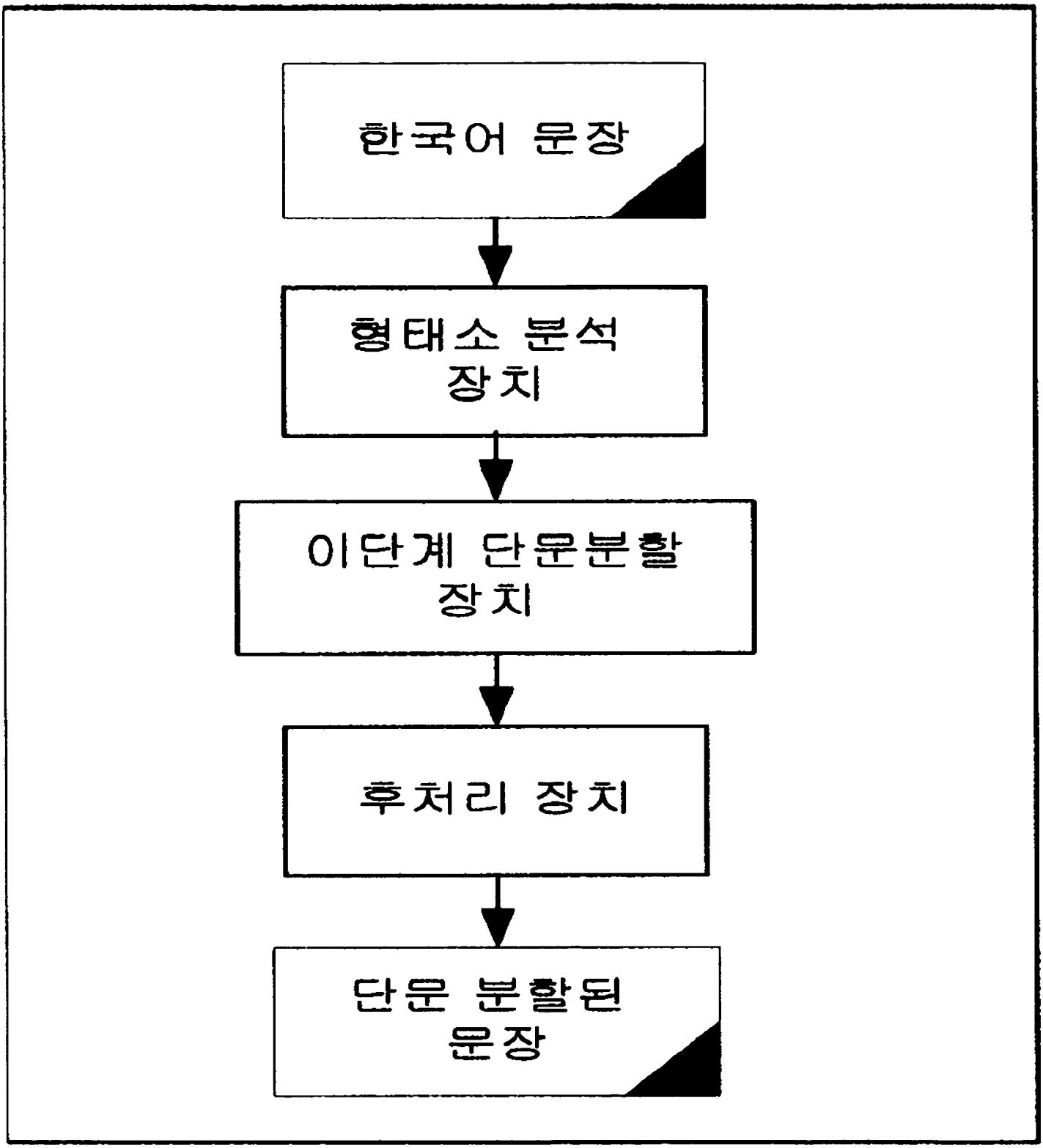
비 핵심성분의 의존 관계 파악과 핵심 구조 범위 확장 통한 의존 구조 확장 장치,

청구항 2.

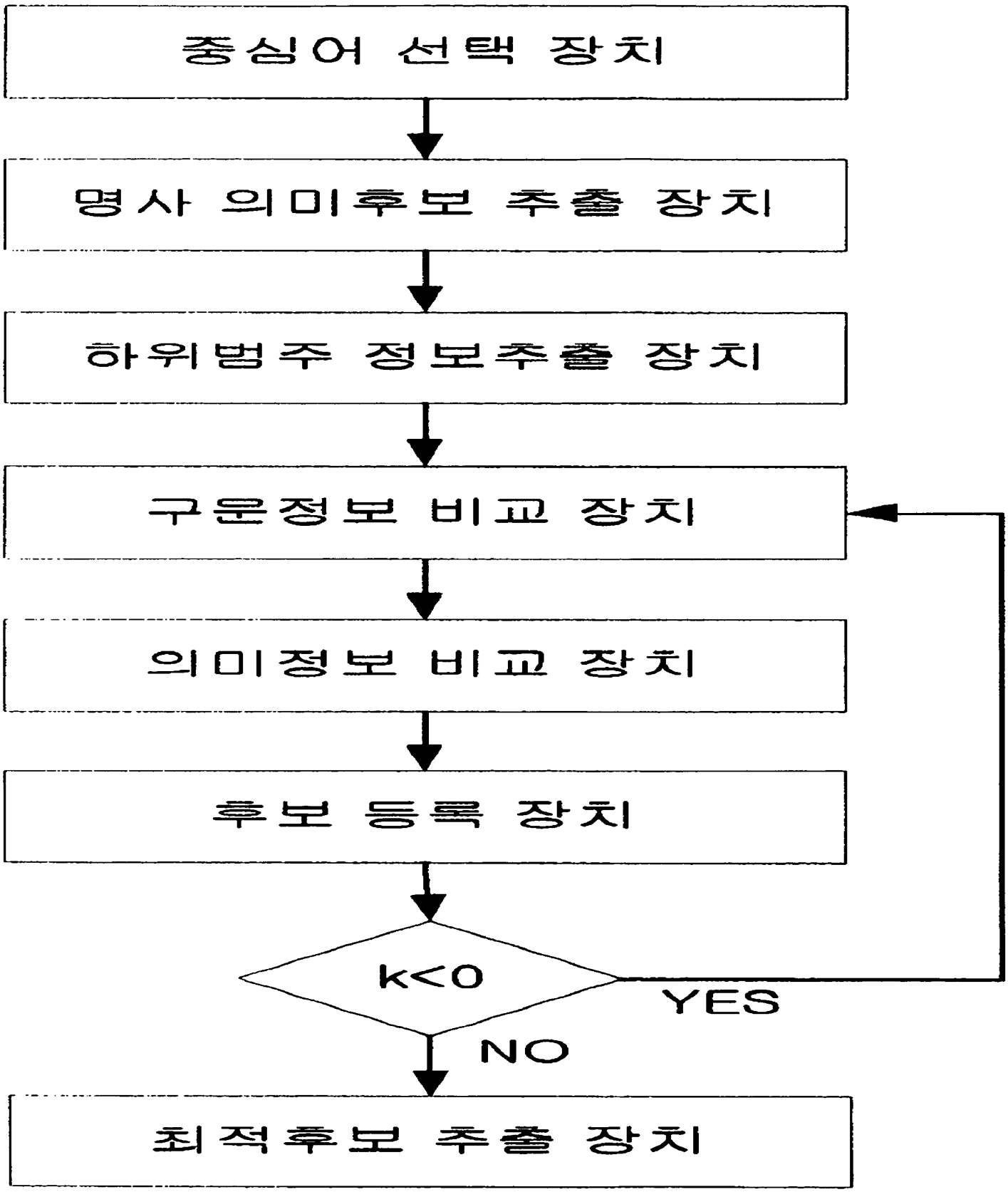
단문 분할 과정에서 구문 정보 비교 장치와 의미 정보 비교 장치를 통한 자동적인 어휘 의미 및 관계성 파악 방법

도면

도면 1



도면 2


중심어 선택 장치
명사 의미후보 추출 장치
하위범주 정보추출 장치
구문정보 비교 장치
의미정보 비교 장치
후보 등록 장치
k<0
YES
NO
최적후보 추출 장치

도면 3

ID	용언	품사	패턴 ID	참고색인			피동 정보	사역 정보	원형 정보	
				1	2	3				

	대표조사				확장조사				의미역				의미마커				예제
	1	2	3	4	1	2	3	4	1	2	3	4	1+	2+	3+	4+	

예1)

1030	몰리다	동사	V6	참고색인	1	0	몰다	
				1_3				

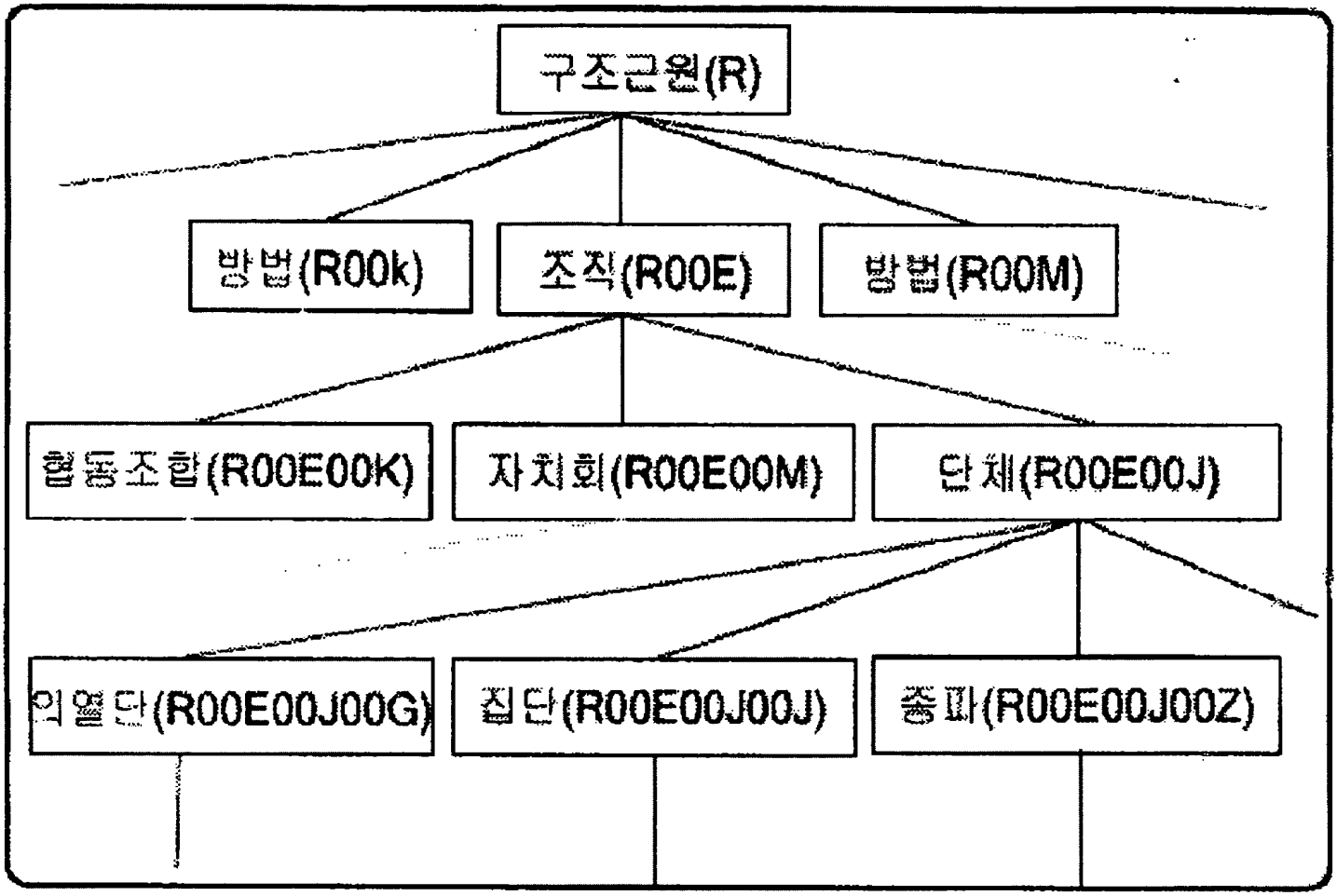
대표조사		확장조사				의미역				의미마커				예제
이	에	이/가/은..	에/까지...	3	4	CHD	GOL	3	4	R00G00B00U01a,...	R00H,..	3+	4+	

예2)

1031	몰리다	동사	V9	참고색인	1	0	몰다	
				1_1				

대표조사		확장조사				의미역				의미마커				예제
이	에(게)서	이/가/은...	로부터/...	3	4	사람	추상물	3	4	1+	R00G00B00U01a	3+	4+	

도면 4


구조근원(R)
방법(R00k)
조직(R00E)
방법(R00M)
협동조합(R00E00K)
자치회(R00E00M)
단체(R00E00J)
연맹단(R00E00J00G)
집단(R00E00J00J)
종파(R00E00J00Z)

도면 5